Award Number: W81XWH-13-1-0020

TITLE: Health-Terrain: Visualizing Large Scale Health Data

PRINCIPAL INVESTIGATOR: Ph.D. Fang, Shiaofen

CONTRACTING ORGANIZATION: Indiana University, Indianapolis, IN 46202

REPORT DATE: April 2015

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
April 2015	Annual	7 MAR 2014 – 6 MAR 2015

4. TITLE AND SUBTITLE
Health-Terrain: Visualizing Large Scale Health Data

5a. CONTRACT NUMBER
W81XWH-13-1-0020

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shiaofen Fang, email: sfang@cs.iupui.edu
Mathew Palakal, Yuni Xia, Shaun J. Grannis, Jennifer L. Williams

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Trustees of Indiana University
980 Indiana Avenue, RM2232
Indianapolis, IN 46202-513-

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
In the past year, we have made significant progress including: (1) creating a concept space data model, which represents a schema tailored to support diverse visualizations and provides a uniform ontology that allows the system to be leveraged for many types of health care datasets through individually designed text and data mining procedures; (2) designing and implementing data and text mining analytics and visualization algorithms; and (3) developing a flexible prototype system for using our analytics and visualization framework to explore large-scale, real-world health data. These three components are integrated in a generalizable browser-based graphical interface, which enables flexible and free-form data exploration and hypothesis discovery and also a more flexible distribution of the resulting software. We have completed the majority of algorithm development and implementation; implementation and testing of a few remaining advanced visualization techniques are outstanding. The system has received favorable initial feedback from users, and we believe it has potential as an open source tool to support health data visualization tasks.

15. SUBJECT TERMS
Health Data, Visualization, large-scale

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	23	

Table of Contents

Introduction

The goal of this project is to develop novel visualization techniques and tools for large and complex health care data to facilitate timely decision-making and trend/pattern detection. A prototype system will be developed to test the effectiveness of this approach on a large-scale health care database that is currently available at Regenstrief Institute. We will develop a public health use case leveraging a Notifiable Condition Detector (NCD) dataset that contains reportable disease conditions that are transmitted to Indiana public health authorities (over 800,000 reports). Clinicians and public health stakeholders seek to uncover informative trends contained within the growing population-based datasets. To support knowledge discovery, in this project, we first extract meaningful terms and their associations and attributes from the raw data by applying data mining and text mining algorithms to construct a concept space. A browser-based user interface will also be developed to enable interactive online data exploration. A suite of visualization algorithms and techniques will be developed and implemented within the prototype system. Visualizations include a novel 3D spatiotemporal terrain visualization technique for big time-series data over the Indiana geographical area.

Body

The primary goals through the aforementioned annual report period are: 1) Design and develop new visualization algorithms and tools; 2) Integrate and test all system functions; and 3) Evaluate system usability.

1. Visualization algorithm design and development

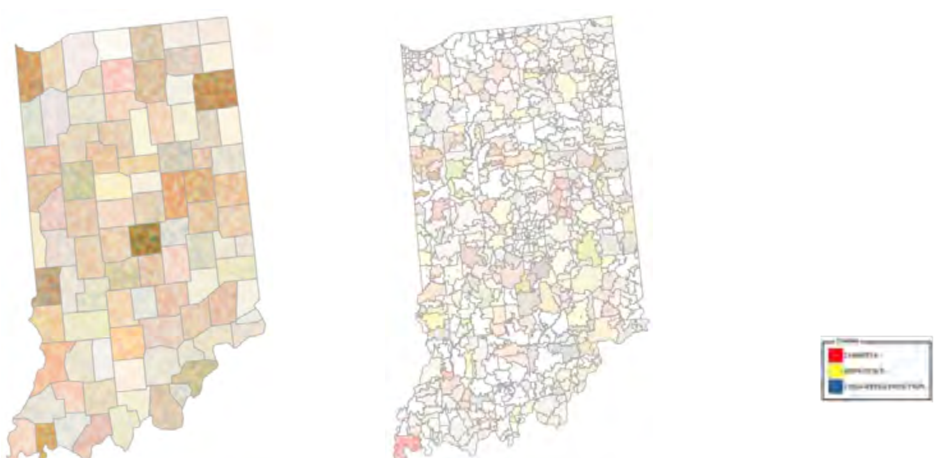
1.1 Spatial Texture Based Approach

Population-level healthcare data and information are often tightly coupled with geospatial regions. The visualization of this type data requires the integration of geo-visualization and multi-dimensional and time-variant information visualization. For this purpose, we propose a Spatial Texture based approach. In this approach, we encode multi-dimensional attributes or time-variant attributes for a geospatial region into a texture image, and then map the texture image to the surface of the geospatial region to provide an integrated visual representation. The key is the visual encoding of multiple attributes or a time-variant attribute in a texture image.

Noise Texture:

We aim to represent multiple attributes for each geospatial region using color coded texture patterns so that the users can visually perceive the representations of different attributes, not only within one region, but also its overall geospatial distributions across many regions in a geographic area (e.g. a state).

We first construct noise patterns to create a random variation in color intensity, similar to the approach in [1]. Different color hues will be used to represent different types of attributes, for example the occurrences of different diseases. A turbulence function [2] will be used to generate the noise patterns of different frequencies (sizes of the sub-regions of the noise pattern). These multi-scale patterns may be applied to different scales of geographic areas (e.g. counties vs zip-codes). Since the noise pattern involves the mixing and blending of different color hues, we choose to use an RYB color model instead of RGB model, as proposed in [1], since RYB color model provides more intuitive representation of the weights of different colors after blending. Figure 1 shows two examples of the texture-mapped views of three diseases, Diabetes. Hepatitis B, and Chlamydia, over the Indiana state map. For example, more reddish areas exhibits higher rate of Diabetes and bluish areas show higher occurrence of Chlamydia.



(a) (b)

Fig. 1 Noise textures mapped over the Indiana State map: (a) county based; (b) zip-code based.

Offset Contours:

Offset contouring is designed to represent attribute changes over time within a geographic region. It can also be used to represent multiple attributes by assigning each attribute to each contour. Similar to the Noise Pattern texture, we first construct a texture image using offset contour curves to form shape-preserving sub-regions. We will then use varying color shades or hues to fill the sequence of sub-regions to represent the change of attribute values over time, or to simply fill the sub-regions with different color values to represent multiple attributes.

The offset contours are generated by offsetting the boundary curve toward the interior of the region, creating multiple offset boundary curves (Figure 2). There are several offset curve algorithms available in curve/surface modeling. But since in our application, the offset curves do not need to be very accurate, we opt to use a simple image erosion algorithm [3] directly on the 2D image of the map to generate the offset contours.

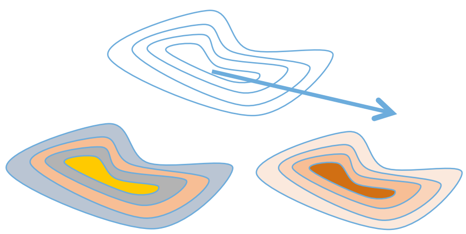
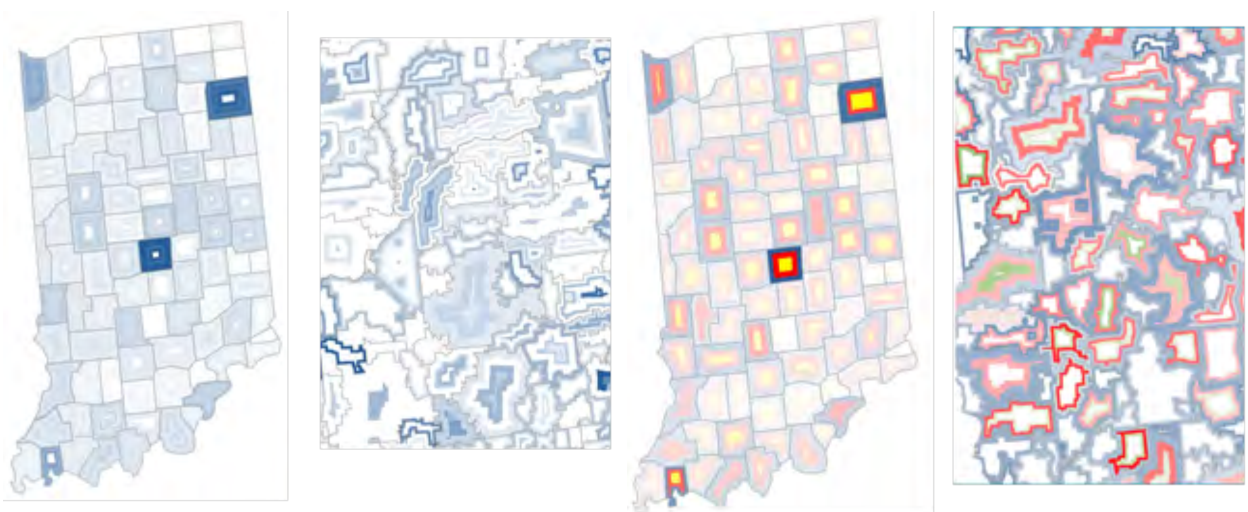


Fig. 2: Offset contours with different colors or different shades of the same color.



(a) (b) (c) (d)

Fig. 3. Texture mapped views of offset contours over the Indiana state map:
(a) County based time-series data
(b) Zip-code based time-series data
(c) County based multi-diseases data
(d) Zip-code based multi-diseases data

In time-series data visualization, the time line can be divided into multiple time intervals and represented by the offset contours. Varying shades of a color hue can be used to represent the attribute changes (e.g. occurrence of a disease) over time. This approach, however, has two limitations. First, when the boundary shape of a region is highly concave, the image erosion technique sometimes does not generate clean offset contours. This usually can be corrected using a geometric offset curve algorithm such as the one in [4]. A second limitation of this approach is that it requires a certain amount of spatial area to layout the contours and color patterns. In public health data, however, these attributes are typically defined on geographic

areas, which provides a perfect platform for texture based visual encoding. Figure 3 shows a few examples of the texture mapped views of offset contours over the Indiana state map. Figure 3 (a-b) show the time-series views of Influenza, from 2004 to 2012. The time interval is divided into 8 subintervals. Figure 3 (c-d) show three diseases, Influenza, Typhoid Fever, and Hepatitis B.

1.2. Spiral Theme Plot

Spatial texture provides overviews of health care data associated with geographic regions. It is however often desirable for health administrators and physicians to also see the details of individual patients and theirs medical history (over time). When this is done with a large population, the collective view of patient medical histories often exhibit identifiable patterns and trends that may not be easily detected from the visualization of statistical data over geographical regions.

A simple approach to view patient level data is to draw each patient record as a point on a radial plot, divided into multiple rings which can represent different terms such as diseases. We call this Ring Plot, as shown in Figure 4. The circumference of this radial space represents the time-axis. Thus, time is encode as the radial angle of the dots (patients). Ring Plot shows the distribution of patient-level data over a time-attribute space. One significant attribute, for example “age”, will be represented as radius. Other attributes of the patients, such as race and gender, are represented as color and shape of the dots. Occurrences of the same patient associated with multiple terms (e.g. diagnosed with multiple diseases) are connected by curves across the graph. In Figure 4, for example, we see a concentration of mid-age patients with Hepatitis B.

Ring Plot, however, does not provide a good overall trend and comparisons of different diseases over time, as typically shown in a ThemeRiver plot. The time axis is also only limited to one circle, which cannot represent periodical patterns very well. By integrating ThemeRiver and a spiral pattern into the basic Ring Plot method, we developed a new Spiral Theme Plot technique. In a Spiral Theme Plot, the diseases (or any other term) are represented as stacked themes along a spiral base curve, which is the time-axis. Patients are still plotted within the regions of the themes, with similar visual features (age, race, gender, etc.). Spiral Theme Plot allows multiple years of patients data be plotted periodically such that seasonal patterns or abnormal patterns for seasonal diseases can be easily detected. For patients with multiple hospital visits at different times for the same or different conditions, curves are drawn to connect these multiple occurrences by the same patient.

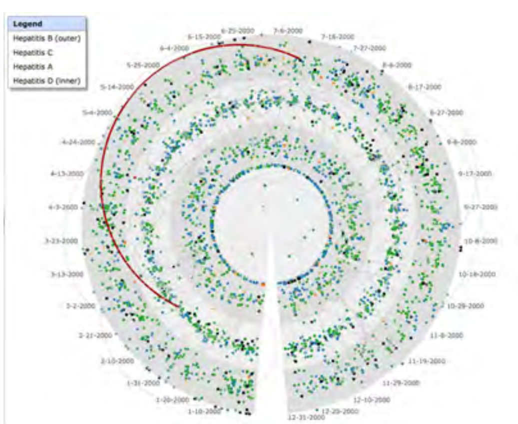
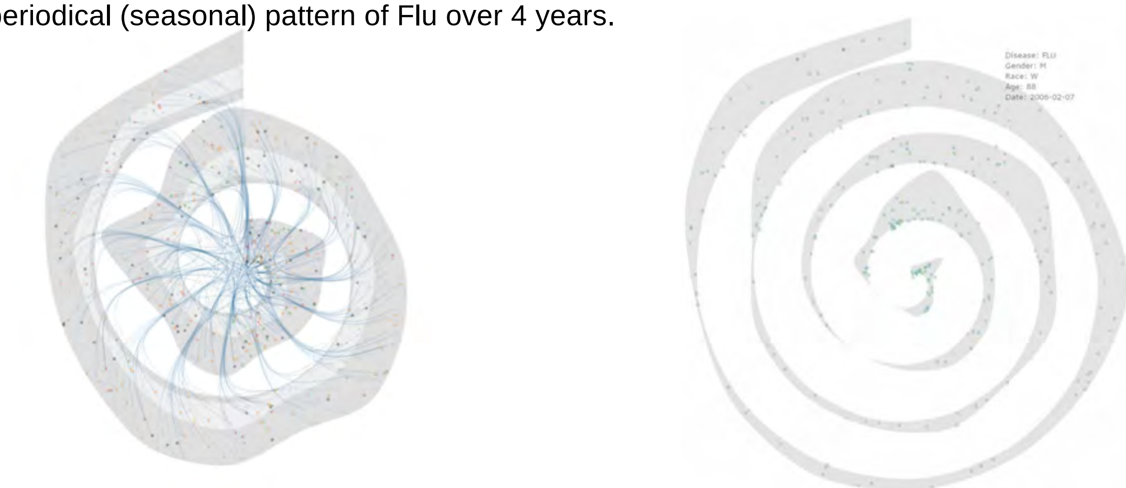


Fig. 4: A Ring Plot for Hepatitis A, B, C and D. For each patient (dot), the color represents race, the shape represents gender, and the radius represents age.

There are several technical details that need to be developed in order to implement a Spiral Theme Plot. First, the radius of the base spiral curve needs to be pre-estimated based on the maximum width of the cumulative themes of all the diseases. Second, when plotting patients within each theme, the width of the theme at that particular angle needs to be computed. Since the boundary curves of the themes are interpolated by spline curves, this width information can theoretically be computed from the spline representations. But we found that it is actually simpler and more efficient to check the color values along the normal direction of the spiral curve to estimate the width of a theme at each angle. Lastly, we found that the number of patients with multiple occurrences is usually quite large, which leads to very dense and cluttered connecting curves. We implemented an edge bundling strategy to bundle these connecting curves for each pre-defined time interval. Other type of bundling strategies may also be implemented to show certain types of connection patterns better. Figure 5a shows an example of a Spiral Theme Plot for Lyme Disease, Blood Diseases, and Brucellosis. Figure 5b show a periodical (seasonal) pattern of Flu over 4 years.



(a) (b)

Fig. 5: Spiral Theme Plot: (a) Lyme Disease, Blood Diseases, and Brucellosis; (b) Seasonal pattern of Flu.

2. System Integration and User Interface

The system is implemented using Javascript in an HTML5 canvas. The architecture pattern is based on the Ruby on Rails (RoR) framework for delivering web applications with AJAX services and a classic Model-View-Controller architecture. The user interface is a modern web GUI using a combination of form submission and RESTful service calls to query and retrieve data in various data delivery formats. The visualization algorithms are implemented using HTML, CSS, SVG, and WebGL technologies with a number of open-source Javascript libraries such as sigma.js, d3.js, jquery.js and three.js.

The user interface uses multiple split windows so that multiple types of visualizations can be applied and compared for the same dataset. Figure 6 show a screen shot of three visualizations for a dataset selected from an association map. Visualization results can also be saved into a slider bar, with time stamps, and be brought back later (Figure 7). This provides a flexible workspace for health administrators or physicians to explore and compare different scenarios for health policy planning, decision making, resource management, etc.

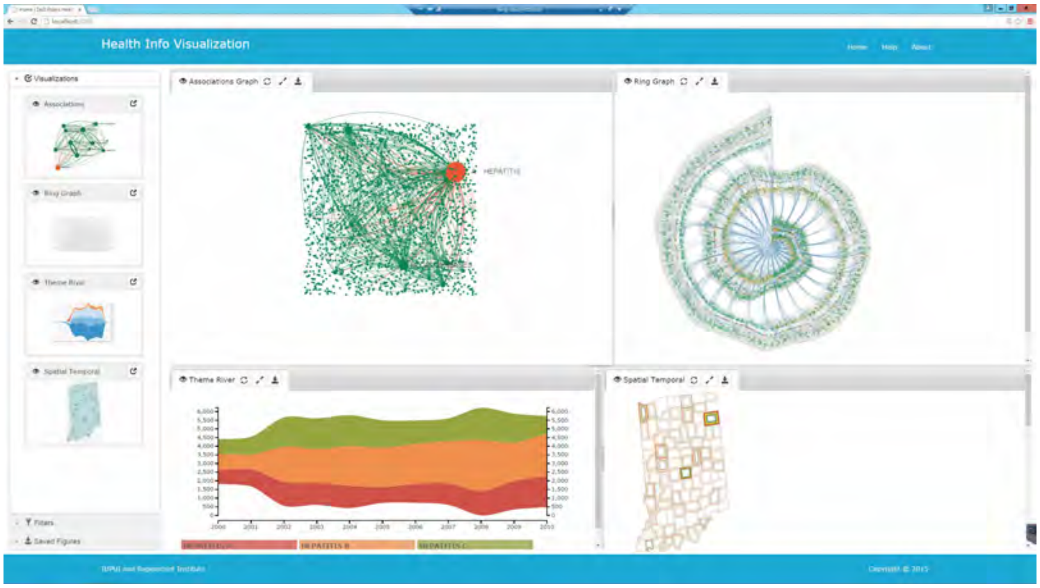


Fig. 6: A screen shot of a split window interface.

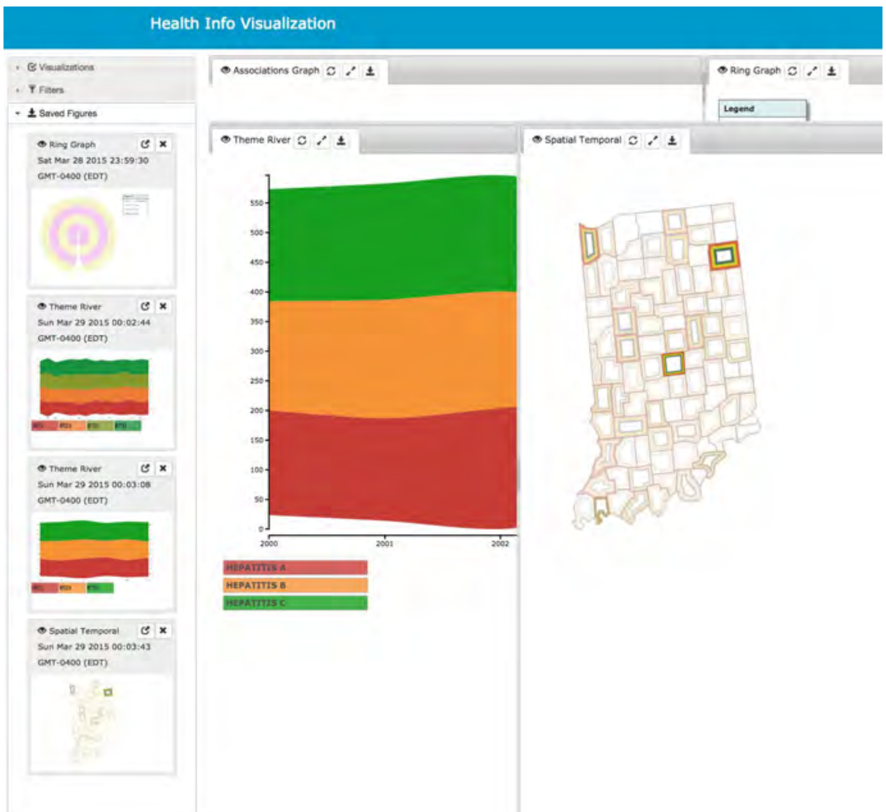


Fig. 7: System interface with saved working windows.

3.   System prototyping and evaluation

After developing the data visualization framework, we imported de-identified communicable disease data and incorporated four operational visualizations including a network association graph, a ring graph, theme river graph, and 2D/3D cholorpleth (heatmap) spatiotemporal graphs. To perform a usability evaluation of this framework we recruited interviewees who represented potential end-users and visualization consumers, including public health epidemiologists with expertise in notifiable disease surveillance and syndromic surveillance; Indiana University

faculty from the school of Public health; biomedical informaticians with public health informatics expertise from the Regenstrief Institute; clinical practitioners; and program managers with advanced training in public health management.

For our usability evaluation we adapted the National Institute of Standards and Technology (2007) definition of usability for our participants as the “effectiveness, efficiency, and satisfaction with which intended users can achieve their tasks and the intended context of product use.” Using an unstructured qualitative interview process, we explored dimensions of effectiveness (accuracy in completing tasks), efficiency (perceived time and effort in accomplishment of tasks), and satisfaction (subjective response to the application).

Prior to reviewing each of the four visualizations in successive order, the interviewees were oriented to the following dimensions of the application: 1) the overall screen layout and structure; 2) the ways in which users could navigate within a screen; 3) the ways in which users could navigate to other screens; 4) the ways in which users could navigate to the home screen; the ways in which users would move from field to field; and 5) a description of key commonly used buttons, icons, and links.

After presenting each visualization, the interviewees were asked to comment on the perceived dimensions of effectiveness, efficiency, and general satisfaction. Where necessary, exemplar leading questions were prepared to stimulate discussion, and included: “comment on your perceived satisfaction with the time required to interact with this visualization”; “how satisfied would you be with the perceived effort to interact with this visualization?”; “how confident are you that you could use this visualization to support your daily work flows on a routine basis?”; and “how quickly do you think most users would learn to perform the functions needed for this visualization?” The interviewees’ responses were synthesized and are summarized below, stratified by each visualization.

Association Network Graph:
As a general theme, the interviewees felt that including in-line guidance or pop-up descriptions (e.g., using mouse-overs) for each visualization parameter would provide end-users with valuable information to guide their use the tool. For example, the purpose of the association "threshold" parameter used in the association graph to create edges was unclear, and interviewees sought further definition. Interviewees noted that the visualization loading time, while less than 10 seconds, could be improved to enhance overall user satisfaction. The meaning of the colors of the edges in the graph was unclear, and interviewees felt they should be more clearly defined in the application. Public health stakeholders expressed the clear value of being able to quickly identify associations among multiple diseases, and they were pleased with the ability to filter out extraneous nodes and create sub-networks for strongly associated diseases. The interviewees felt that edges in the graph should contain metrics characterizing the strength of the association between nodes (disease).

Ring Graph:
The interviewees described this visualization as being particularly complex and exhibiting high information density; some felt that the density obfuscated important information and were concerned that individual cases may be overlooked. The interviewees required substantial introduction to the graph prior to expressing recognition of the value of the visualization. Several commented that the extended (90-120 second) loading time was sub optimal, and hindered overall satisfaction, usability, and efficiency. While the dimensionalities of disease, age, gender, race, and time were generally perceived to be useful, the interviewees suggested that allowing those dimensions to be configurable would improve the utility of this visualization. One epidemiologist interviewee noted that their team likely would not use this visualization to identify disease outbreaks, but would instead use this visualization after an outbreak has been detected through other means in order to explore the relationships and characteristics of individuals within an outbreak in order to identify potential risk factors and target interventions. Another

suggested that the circular format could be confusing and may obfuscate data; it was suggested that the graph be transformed into a linear format to potentially improve interpretability. One interviewee noted, “This graph has the potential to make me think about things that I wouldn't otherwise, and that has value to me.”

Theme River Graph:
Interviewees generally expressed that the theme river visualization provided a consumable, informative high-level comparison communicable disease incidence over time. Multiple interviewees indicated they would prefer case counts to begin at a common baseline on the y-axis; the variable heights and irregular sides of the theme river graph were felt to hinder interpretability. A consistent linear y-axis baseline of zero was felt to potentially enhance year-to-year comparisons over the default theme River visualization.

Spatiotemporal Graph:
The three-dimensional version of this visualization was perceived to be more informative than the two-dimensional version. Commenters noted that the two-dimensional color variations within counties were challenging to interpret; the varying color intensity combined with varying band widths for each disease confused the interviewees. Some noted that continuous variation in color intensity may be less interpretable than dividing the range of disease incidence into a discrete set of ranges. Interviewees stated that presenting disease incidence as a three-dimensional height substantially improved interpretability and understanding of the data. There is wide variation in disease rates among counties (a small number of counties contain significant portions of overall disease); this variation obfuscates details in lower prevalence regions. Consequently the interviewees suggested that an additional feature enabling nonlinear scaling to highlight details in lower prevalence counties would be useful. They further suggested that presenting these data as incident rates (new cases per total population in the county) versus absolute counts (new cases) could improve interpretability and overall satisfaction. Epidemiologist interviewees requested extended functionality to visualize the highest prevalence diseases in each county.

General observations:
Due to the data privacy policy provisions of the institutional review board research process, we used obfuscated de-identified clinical data for the usability assessment. The interviewees noted that further assessment of the usability of these different visualization tools would be enhanced by reviewing fully identified data rather than the de-identified obfuscated data currently used for research and development purposes.

Key Research Accomplishments

- Designed and implemented several information visualization algorithms suitable for public health data, including the texture based geospatial and geo-temporal visualization techniques, and patient based plotting methods: Ring graph and Spiral Theme Plot.
- Designed and implemented a web based graphical user interface for the prototype system, including a multi-window interface
- Completed demonstrations to public health staff/providers and biomedical informaticians for the system evaluation.

Reportable Outcomes

- S. Jiang, S. Fang, S. Bloomquist, J. Keiper, M. Palakal, Y, Xia, S. Grannis. Health-Terrain: A Healthcare Data Visualization System. Submitted: 2015 IEEE Symposium on Large Data Analytics and Visualization.
- Three demonstrations of the system were completed within this timeframe; one for the Indiana State Department of Public Health, one for the Marion County Public Health Department, and one for the internal departments of IUPUI School of Informatics and Regenstrief Institute - Indiana University School of Medicine. Feedback from the aforementioned demonstrations will be utilized to further enhance the system.

Conclusion

We have made significant overall progress in this project, including: (1) Design and development of new visualization algorithms and tools; (2) Integration and testing of all system functions; and (3) Evaluation of system usability. We focus on two new visualization methods we developed specifically for public health data: Spatial Textures, and Spiral Theme Plot. Spatial Texture approach is effective because geospatial visualization intrinsically provides additional screen space (surface areas) that can be taken advantages of to encode additional data and attributes. The Spiral Theme Plot technique is a combination of several information visualization methods including ThemeRiver, Spiral Plot and Scatter Plot. For public health data with large patient databases, this particular combination satisfies several key requirements for visualizing time-variant patient records. With the rich set of tools available to support web based user interface, graphics, and data communications, we also feel that it is as efficient to develop a web based visualization system as in a traditional programming environment. We are currently carrying out integrated system testing and evaluation, and we are confident that an innovative and easy to use health care data visualization system will be available by the end of September 2015.

References

[1] Nathan Gossett, Baoquan Chen. Paint Inspired Color Mixing and Compositing for Visualization. IEEE Symposium on Information Visualization 2004. 113-117.

[2] Ken Perlin. An image synthesizer. In Proceedings of SIGGRAPH85, pages 287–296. ACM Press, 1985.

[3] Rosenfeld, A. and A.C. Kak (1982). Digital Picture Processing. Academic Press, New York.

[4] Hoschek, J., (1988), “Spline Approximation of Offset Curves,” Computer Aided Geometric Design, Vol. 5, pp. 33–40.

Appendices

Health-Terrain: A Healthcare Data Visualization System

Abstract—Healthcare data visualization is challenging due to the needs for integrating geospatial information, temporal information, text information, and heterogenious health attributes within a common visual context. We recently developed a web-based healthcare data visualization system, Health-Terrain, based on a Notifiable Condition Detector (NCD) use case. In this paper, we will describe this Helath-Terrain system, with emphasis on the visualization techniques developed specifically for healthcare data. Two new visualization techniques will be described: (1) A spatial texture based visualization approach for multi-dimensional attributes and time-series data; (2) A spiral theme plot technique for visualizing time-variant patient data.

Index Terms— healthcare data, spatiotemporal visualization, geospatial information visualization, data and text mining, web-based visualization systems

1 Introduction

As electronic healthcare systems are being fully integrated nationally, the effective visualization of large and complex healthcare data becomes increasingly desirable for timely decision making and trend/pattern detection [1]. The problem, however, is very challenging for several reasons:

1) Health data is a data-rich, information-poor domain. In Electronic Health Record (EHR) systems, data are almost always heterogeneous, unstructured, hierarchical, and longitudinal.
2) EHR systems are often extremely large. While it is possible to visualize an EHR system in small scales and with a focused scope, high impact knowledge discoveries more likely come from global scale (population-wide) visualization and knowledge mining.
3) Visualizing population-level health data often involves presenting geospatial and time-series data in a common visual context. This presents a challenge in visual encoding of the information space.

For heterogeneous and complex data, feature extraction through data mining is critical, as visualizing a feature space is much more feasible. For healthcare data, this feature space often consists of healthcare terms (ontology) and their relationships. Therefore, the effective integration of data processing, data mining, and text mining is necessary in healthcare data visualization. Although healthcare data is very large, the visualization of aggregated features, combined with patient level visualization, can be very effective in revealing the patterns and trends of population health. It is therefore important to develop multiple visualization tools to be integrated within a common visual interface to allow users to visually explore the data through an easily accessible platform such as a web browser.

One of the unique challenges in healthcare data visualization is how to visualize multi-attributes and time-series data with associated geospatial information. In our approach, we embed multiple attributes and the time variable within a geospatial representation to take advantage of the available geographic space. This can be done by mapping texture images onto the geospatial surfaces. The key is then to properly represent the multi-attributes and time-series information in a texture image by constructing visually effective texture representations. While visualizing aggregated data for geospatial areas provides global trends and patterns in a geospatial context, we are often interested in visualizing individual patient records and their development over time. To this end, we also developed a spiral theme plot technique for visualizing time-variant patient records and attributes. These new visualization techniques have been implemented in a web-based healthcare data visualization system called Health-Terrain, and tested on real healthcare databases.

In the rest of this paper, we first discuss some related work in Section 2, and then describe the overall functionalities and data processing and mining modules in Section 3. Section 4 and Section 5 will focus on the two new visualization techniques: spatial textures and spiral theme plots. Section 6 will provide some implementation details and some additional visualization results. We conclude the paper in Section 7.

2 Related Work

The visualization of large scale healthcare data has not been extensively studied. There are several existing works and visualization systems that deal with the secondary use of electronic health record data in a limited scope. LifeLines [2] uses a traditional 2D time line visualization technique to visualize specific patient medical and health history. It emphasizes the visualization of temporal ordering of events with limited aggregation effect. An extension of LifeLine, LifeLine2 [3], enables multiple patient comparisons and aggregation for analysis, but the visualization design limited its scalability. A similar system, call TimeLine [4], re-organizes and re-groups multiple EHR content types in a layout of Y-axis to track multiple events along the same time line. A set of visualization tools are described for visualizing a patient's electronic health record to aid physicians' diagnosis and decision-making. The traditional matrix view and parallel coordinates are the main techniques applied. The VISITORS system [5,6] combines a clinical knowledge base with visualization to enable users to explore multiple clinical records. It relies on domain ontologies to define clinically meaningful higher abstractions given raw, temporal data. CLEF [7] is a system enabling visual navigation through a patient's medical record using semantically and temporally organized networks to represent events throughout the patient's medical history. CLEF also supports limited text processing capabilities for generating textual summaries. None of these existing systems is capable of visualizing large-scale integrated EHR datasets. A review paper on visualization tools for infectious diseases is given in [8].

Population-level healthcare data visualization involves both geospatial information and time-variant attributes. The geospatial visualization of time-series data is challenging because it is difficult to encode the time axis in a geospatial context. Animation based techniques (e.g. [9]) do not provide a good space-time overview. Other techniques, such as color-coding of time [10], connecting time-lines [11], and time-curves [12], often introduces visual clutter and occlusion, which are infeasible for large scale datasets. A well-known technique in geospatial time-series visualization is Space-Time-Cube [13-17]. It is a 3D representation of a combination of time axis (Z-axis) and a 2D geographic map (X-Y plane). Time-lines or time-curves are used to depict data evolution over time. While time and spatial information are integrated in a 3D visual representation in a space-time-cube, the sense of space-time embedding diminishes as the data moves up in the time axis. Visual clutter will also be a problem with large datasets. Many other techniques have been developed for the visualization of time-series data without explicit geospatial information such as time-series plot

[18] and ThemeRiver [19]. Many variations of ThemeRiver styled techniques have been applied in different time-series visualization applications, in particular text visualization [20]. Spiral patterns have also been used in visualizing time-series data [21] to provide better identification of periodic structures in the data.

Texture-based visualization techniques have been widely used for vector field data, in particular, flow visualization. Typically, a grayscale texture is smeared in the direction of the vector field by a convolution filter, for example, the Line Integral Convolution (LIC), such that the texture reflects the properties of the vector field [22,23,24]. Similar techniques have also been applied to tensor fields [25,26].

3 The Health-Terrain System

3.1 System Overview and Use Case

Our goal is to develop a prototype system, Health-Terrain, to support visual exploration of large healthcare data sets on a browser based interface. The system integrates information visualization, web-based user interaction, and text and data mining techniques. A concept space approach is used to unify data representation unified data representation through data and text mining. Figure 1 shows the system architecture with all three aspects of the system components: interface, algorithms and data models.

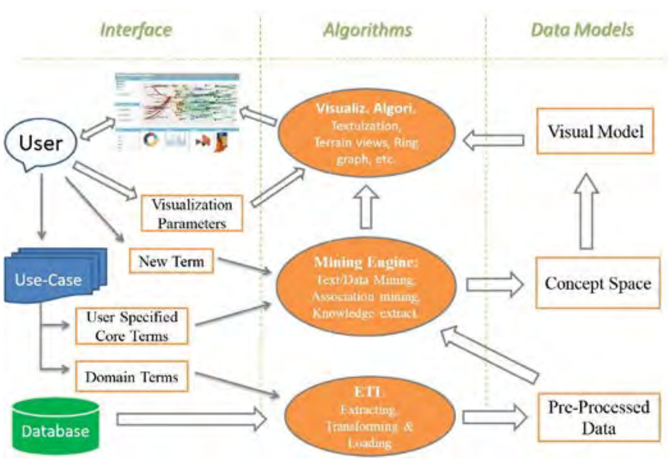


Fig. 1. Health-Terrain system architecture and components.

To test our visualization system we used a large public health notifiable disease reporting system. The XXXX Institute implemented and maintains an unparalleled HIE-based, automated electronic lab reporting (ELR) and case-notification system for over ten years in the State of XXXX. The Notifiable Condition Detector (NCD) System uses a standards-based messaging and vocabulary infrastructure that includes Health Level Seven (HL7) and Logical Observation Identifiers Names and Codes (LOINC) [27]. The NCD receives real-time HL7 version 2 clinical transactions daily, including diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations. The system automatically detects positive cases of notifiable conditions and forwards alerts to local and state health departments for review and follow up. These alerts enable more effective and efficient public health population health monitoring and case management. The NCD dataset contains 833,710 public health notifiable cases spanning more than 10 years from among 439,547 unique patients. An additional dataset containing 325,791 unstructured clinical discharge summaries, laboratory reports, and patient histories were extracted. In order to comply with the patient privacy policies and protocols of the institutes where the datasets came from, the actual data visualized in this paper has been altered or perturbed. Nevertheless, the overall patterns and trends of the data are generally preserved.

3.2 Concept Space

The "concept space" represents a uniform layer of clinical observations and their associations, and enables users to explore data using various visualization and analysis methods. Concept terms are derived from data mining and text-mining processes applied to the use case datasets. Disease concepts were extracted from the NCD dataset. Text mining algorithms were then applied to additional linked text dataset (unstructured clinical summaries) to construct ontologies for different concept types, including disease, symptom, mental behaviour, and risky behaviour.

The concept space uses a controlled vocabulary that can be pre-defined based on application needs, and enhanced by data/text mining algorithms. These terms and their relationships are represented in an association map, as a space of extracted partial knowledge. This association map is often the starting point of a visual exploration process. Figure 2 shows an example of the association map of diseases. Association map is a graph visualization of the association relationships among the diseases and other terms in the concept space. It can serve as a platform supporting interactive selection of concepts to dynamically visualize data using a variety of tools in the visualization system. To draw an association graph, a spring-embedder algorithm [28] is used to layout the graph nodes. Nodes picked on the association map are then be visualized with geospatial information, possibly with time varying variables. .

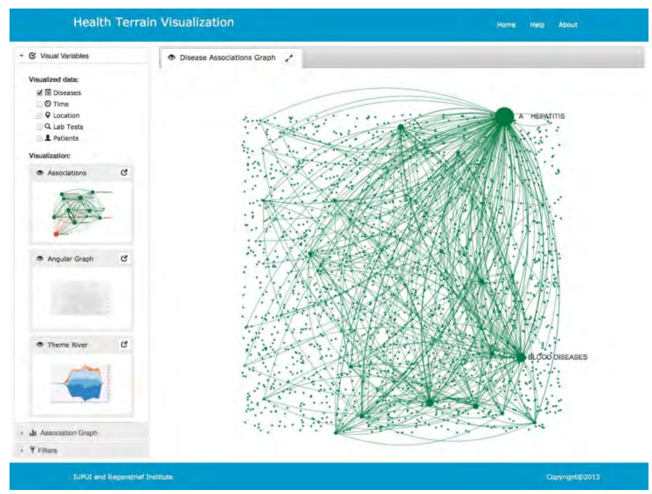


Fig. 2. Disease association map.

In text mining, we processed 325,791 unstructured clinical notes containing patient discharge summaries, laboratory reports, and medical histories. Advanced NLP was applied in the form of named entity recognition (NER) for extracting diseases and other terms, with the help of the Unified Medical Language System (UMLS) [29]. Stemming and concept clustering algorithms [30] were applied to normalize the lexical variants and duplications of the terms. Term correlations were computed using the tf-idf (term frequency – inverse document frequency) vector space model to identify the significantly co-occurring diseases. An association-mining algorithm was applied to the combined terms to generate an association graph among all the concepts terms. The resulting concept space, along with the processed NCD data, is represented in a data model designed to support our specific ontology.

4 Spatial Texture Based Approach

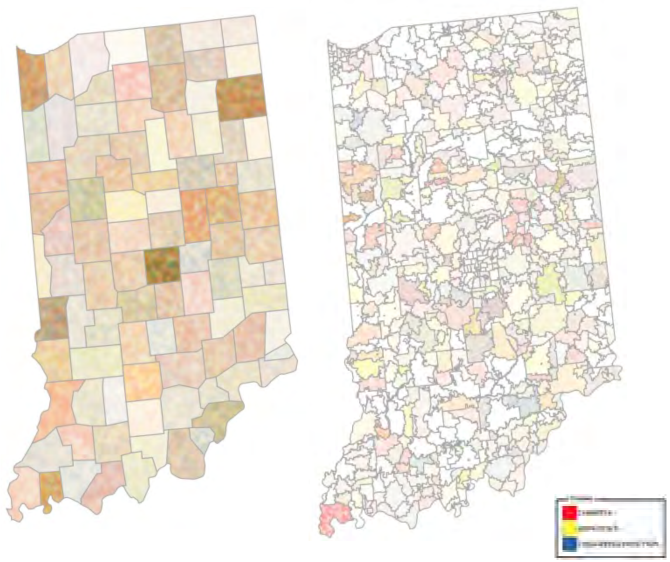
Population-level healthcare data and information are often tightly coupled with geospatial regions. The visualization of this type data requires the integration of geo-visualization and multi-dimensional and time-variant information visualization. For this purpose, we propose a Spatial Texture based approach. In this approach, we encode multi-dimensional attributes or time-variant attributes for a

geospatial region into a texture image, and then map the texture image to the surface of the geospatial region to provide an integrated visual representation. The key is the visual encoding of multiple attributes or a time-variant attribute in a texture image.

4.1 Noise Texture

We aim to represent multiple attributes for each geospatial region using color coded texture patterns so that the users can visually perceive the representations of different attributes, not only within one region, but also its overall geospatial distributions across many regions in a geographic area (e.g. a state).

We first construct noise patterns to create a random variation in color intensity, similar to the approach in [31]. Different color hues will be used to represent different types of attributes, for example the occurrences of different diseases. A turbulence function [32] will be used to generate the noise patterns of different frequencies (sizes of the sub-regions of the noise pattern). These multi-scale patterns may be applied to different scales of geographic areas (e.g. counties vs zip-codes). Since the noise pattern involves the mixing and blending of different color hues, we choose to use an RYB color model instead of RGB model, as proposed in [31], since RYB color model provides more intuitive representation of the weights of different colors after blending. Figure 3 shows two examples of the texture mapped views of three diseases, Diabetes. Hepatitis B, and Chlamydia, over the Indiana state map. For example, more reddish areas exhibits higher rate of Diabetes and bluish areas show higher occurrence of Chlamydia.



(a) (b)

Fig. 3 Noise textures mapped over the Indiana State map: (a) county based; (b) zip-code based.

4.2 Offset Contours

Offset contouring is designed to represent attribute changes over time within a geographic region. It can also be used to represent multiple attributes by assigning each attribute to each contour. Similar to the Noise Pattern texture, we first construct a texture image using offset contour curves to form shape-preserving sub-regions. We will then use varying color shades or hues to fill the sequence of sub-regions to represent the change of attribute values over time, or to simply fill the sub-regions with different color values to represent multiple attributes.

The offset contours are generated by offsetting the boundary curve toward the interior of the region, creating multiple offset boundary curves (Figure 4). There are several offset curve algorithms available in curve/surface modeling. But since in our application, the offset curves do not need to be very accurate, we opt to use a simple image erosion algorithm [33] directly on the 2D image of the map to generate the offset contours.

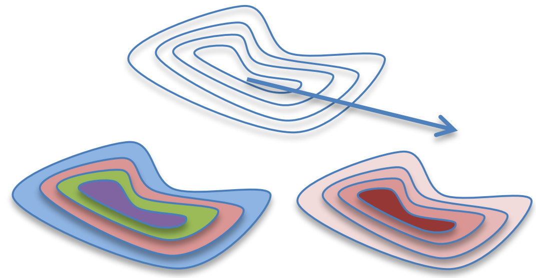
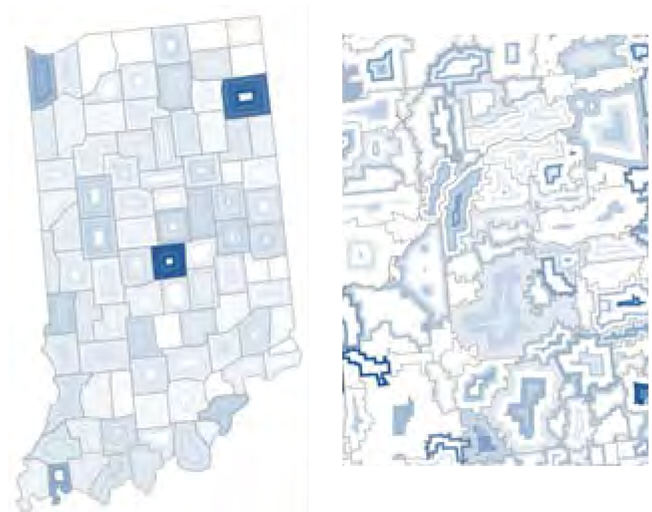
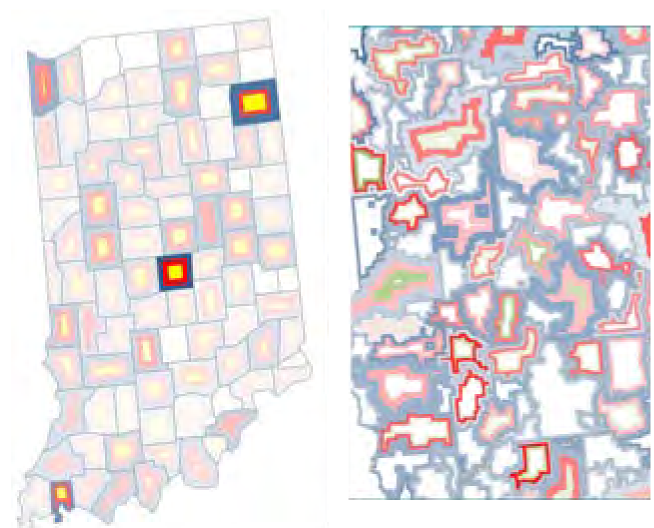


Fig. 4: Offset contours with different colors or different shades of the same color.



(a) (b)



(c) (d)

Fig. 5. Texture mapped views of offset contours over the Indiana state map: (a) County based time-series data; (b) Zip-code based time-series data; (c) County based multi-diseases data; (d) Zip-code based multi-diseases data.

In time-series data visualization, the time line can be divided into multiple time intervals and represented by the offset contours. Varying shades of a color hue can be used to represent the attribute changes (e.g. occurrence of a disease) over time. This approach,

however, has two limitations. First, when the boundary shape of a region is highly concave, the image erosion technique sometimes does not generate clean offset contours. This usually can be corrected using a geometric offset curve algorithm such as the one in [34]. A second limitation of this approach is that it requires a certain amount of spatial area to layout the contours and color patterns. In public health data, however, these attributes are typically defined on geographic areas, which provides a perfect platform for texture based visual encoding. Figure 5 shows a few examples of the texture mapped views of offset contours over the Indiana state map. Figure 5 (a-b) show the time-series views of Influenza, from 2004 to 2012. The time interval is divided into 8 subintervals. Figure 5 (c-d) show three diseases, Influenza, Typhoid Fever, and Hepatitis B.

5 SPIRAL THEME PLOT

Spatial texture provides overviews of health care data associated with geographic regions. It is however often desirable for health administrators and physicians to also see the details of individual patients and theirs medical history (over time). When this is done with a large population, the collective view of patient medical histories often exhibit identifiable patterns and trends that may not be easily detected from the visualization of statistical data over geographical regions.

A simple approach to view patient level data is to draw each patient record as a point on a radial plot, divided into multiple rings which can represent different terms such as diseases. We call this Ring Plot, as shown in Figure 6. The circumference of this radial space represents the time-axis. Thus, time is encode as the radial angle of the dots (patients). Ring Plot shows the distribution of patient-level data over a time-attribute space. One significant attribute, for example "age", will be represented as radius. Other attributes of the patients, such as race and gender, are represented as color and shape of the dots. Occurrences of the same patient associated with multiple terms (e.g. diagnosed with multiple diseases) are connected by curves across the graph. In Figure 6, for example, we see a concentration of mid-age patients with Hepatitis B.

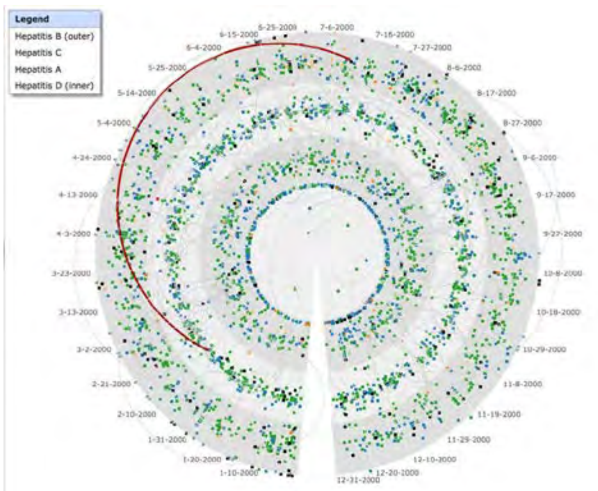


Fig. 6: A Ring Plot for Hepatitis A, B, C and D. For each patient (dot), the color represents race, the shape represents gender, and the radius represents age.

Ring Plot, however, does not provide a good overall trend and comparisons of different diseases over time, as typically shown in a ThemeRiver plot. The time axis is also only limited to one circle, which cannot represent periodical patterns very well. By integrating ThemeRiver and a spiral pattern into the basic Ring Plot method, we developed a new Spiral Theme Plot technique. In a Spiral Theme Plot, the diseases (or any other term) are represented as stacked themes along a spiral base curve, which is the time-axis. Patients are still plotted within the regions of the themes, with similar visual features (age, race, gender, etc.). Spiral Theme Plot allows multiple years of patients data be plotted periodically such that seasonal patterns or abnormal patterns for seasonal diseases can be easily detected. For patients with multiple hospital visits at different times for the same or different conditions, curves are drawn to connect these multiple occurrences by the same patient.

There are several technical details that need to be developed in order to implement a Spiral Theme Plot. First, the radius of the base spiral curve needs to be pre-estimated based on the maximum width of the cumulative themes of all the diseases. Second, when plotting patients within each theme, the width of the theme at that particular angle needs to be computed. Since the boundary curves of the themes are interpolated by spline curves, this width information can theoretically be computed from the spline representations. But we found that it is actually simpler and more efficient to check the color values along the normal direction of the spiral curve to estimate the width of a theme at each angle. Lastly, we found that the number of patients with multiple occurrences is usually quite large, which leads to very dense and cluttered connecting curves. We implemented an edge bundling strategy to bundle these connecting curves for each pre-defined time interval. Other type of bundling strategies may also be implemented to show certain types of connection patterns better. Figure 7 shows an example of a Spiral Theme Plot for Lyme Disease, Blood Diseases, and Brucellosis. Figure 8 show a periodical (seasonal) pattern of Flu over 4 years.

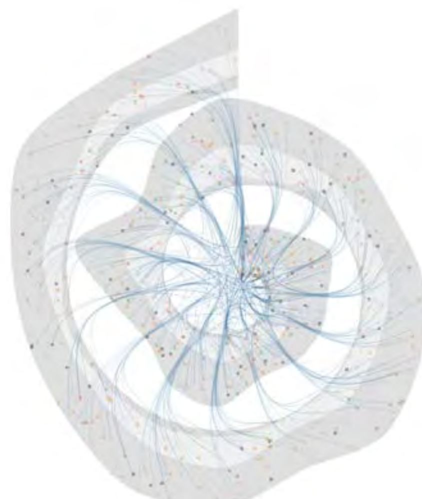


Fig. 7: A Spiral Theme Plot for Lyme Disease, Blood Diseases, and Brucellosis.

Fig. 8: Seasonal pattern of Flu.

6 System Implementation and Interface

The system is implemented using Javascript in an HTML5 canvas. The architecture pattern is based on the Ruby on Rails (RoR) framework for delivering web applications with AJAX services and a classic Model-View-Controller architecture. The user interface is a modern web GUI using a combination of form submission and RESTful service calls to query and retrieve data in various data delivery formats. The visualization algorithms are implemented using HTML, CSS, SVG, and WebGL technologies with a number of open-source Javascript libraries such as sigma.js, d3.js, jquery.js and three.js.

The user interface uses multiple split windows so that multiple types of visualizations can be applied and compared for the same dataset. Figure 9 show a screen shot of three visualizations for a dataset selected from an association map. Visualization results can also be saved into a slider bar, with time stamps, and be brought back later (Figure 10). This provides a flexible workspace for health administrators or physicians to explore and compare different scenarios for health policy planning, decision making, resource management, etc.

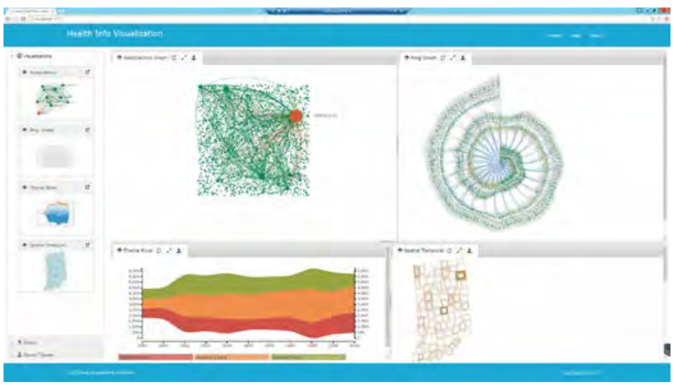


Fig. 9: A screen shot of a split window interface.

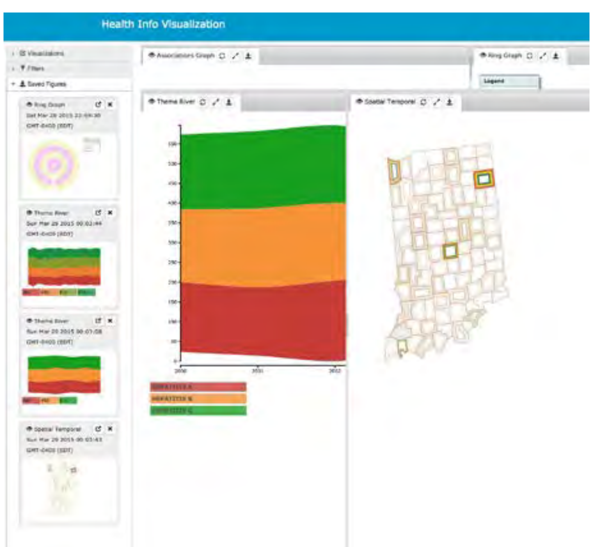


Fig. 10: System interface with saved working windows.

7 Conclusions

We present a visualization system for large healthcare data. We focus on two new visualization methods we developed specifically for public health data: Spatial Textures, and Spiral Theme Plot. Spatial Texture approach is effective because geospatial visualization intrinsically provides additional screen space (surface areas) that can be taken advantages of to encode additional data and attributes. The Spiral Theme Plot technique is a combination of several information visualization methods including ThemeRiver, Spiral Plot and Scatter Plot. For public health data with large patient databases, this particular combination satisfies several key requirements for visualizing time-variant patient records. With the rich set of tools available to support web based user interface, graphics, and data communications, we also feel that it is as efficient to develop a web based visualization system as in a traditional programming environment.

In the future, we would like to continue refining and expanding this visualization system by adding new visualization tools and improving the existing tools. We would also like to develop a configurable user and data interface so that the system can be easily configured for other types of use cases in public health applications.

References

[1] Grossman C, Powers B, McGinnis JM (Ed). Digital infrastructure for the learning health care system: the foundation for continuous improvement in health and health care. The National Academies Press, 2011

[2] Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B., Lifeline: Visualizing Personal Histories, CHI, 1996, pp. 221-227.

[3] Wang, T.D., Plaisant, C., Quinn, A.J., Stanchak, R., Murphy, S., Shneiderman, B. Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records, CHI'08, 2008, pp. 457-466.

[4] Bui, A., Aberle, D.R., Kangarloo, H. Timeline: Visualizing Integrated Patient Records. IEEE Trans. Information Technology in Biomedicine 11(4):462-473.

[5] Klimov D, Shahar Y. A framework for intelligent visualization of multiple time-oriented medical records. AMIA Annual Symp Proc. 2005; 2005:405–409

[6] Klimov D, Shahar Y, Taieb-Maimon M. Intelligent interactive visual exploration of temporal associations among multiple time-oriented patient records. Methods Inf Med. 2009;48: 254–262

[7] Hallett, C. Multi-Modal Presentation of Medical Histories. IUI'08: 13th International Conference on Intelligent User Interfaces. 2008, pp. 80-89.

[8] Carroll LN et al. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.04.006

[9] GEMMELL J., ARIS A., LUEDER R.: Telling stories with MyLifeBits. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on (2005), IEEE, pp. 1536–1539.

[10] THE NEW YORK TIMES COMPANY: Openpaths, Feb. 2013. URL: https://openpaths.cc.

[11] GOOGLE: Latitude, Feb. 2013. URL: http://www.google.com/latitude/.

[12] ECCLES R., KAPLER T., HARPER R., WRIGHT W.: Stories in GeoTime. In VAST (Oct. 2007), Ieee, pp. 19–26.

[13] M. Kraak, "The Space Time Cube Revisited from a Geovisualization Perspective," Proc. 21st Int'l Cartographic Conf., pp. 1988-1996, 2003.

[14] Kraak, Menno-Jan, and P. F. Madzudzo. "Space time visualization for epidemiological research." ICC 2007: Proceedings of the 23nd international cartographic conference ICC: Cartography for everyone and for you. 2007.

[15] Kraak, M. J. and A. Kousoulakou (2004). A visualization environment for the space-time-cube. Developments in spatial data handling 11th International Symposium on Spatial Data Handling. P. F. Fisher. Berlin, Springer Verlag: 189-200.

[16] Kwan, M. P. (2000). "Interactive geovisualization of activity travel patterns using three-dimensional geographical information systems: a

methodological exploration with a large data set." Transportation Research C 8: 185-203

[17] Andrienko, N., G. L. Andrienko, et al. (2003). Visual data exploration using space-time cube. 21st International Cartographic Conference, Durban, South Africa.

[18] Edward R. Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 1983

[19] Havre, S., Richland, WA, Hetzler, B., ThemeRiver: visualizing theme changes over time. IEEE Information Visualization, 2000. 115-123.

[20] W Cui, S Liu, L Tan, C Shi, Y Song, Z Gao, H Qu, X Tong. Textflow: Towards better understanding of evolving topics in text. IEEE Transactions on Visualization and Computer Graphics, 17 (12), 2412-2421

[21] Marc Weber , Marc Alexa , Wolfgang Müller. Visualizing Time-Series on Spirals. IEEE Information Visualization, 2001, 7-13.

[22] B. Cabral and L. C. Leedom. Imaging Vector Fields Using Line Integral Convolution. In Poceedings of ACM SIGGRAPH 1993, Annual Conference Series, pages 263–272, 1993

[23] D. Stalling and H. Hege. Fast and Resolution Independent Line Integral Convolution. In Proceedings of ACM SIGGRAPH 95, Annual Conference Series, pages 249–256. ACM SIGGRAPH, 1995.

[24] R. S. Laramee, H. Hauser, H. Doleisch, F. H. Post, B. Vrolijk, and D. Weiskopf. The State of the Art in Flow Visualization: Dense and Texture-Based Techniques. Computer Graphics Forum, 3(2):203–221, June 2004.

[25] T. McGraw, M.Nadar. Fast Texture-Based Tensor Field Visualization for DT-MRI. 4th IEEE International Symposium on Biomedical Imaging: Macro to Nano, pp. 760-763, 2007.

[26] Cornelia Auer, Claudia Stripf, Andrea Kratz, Ingrid Hotz. Glyph- and Texture-based Visualization of Segmented Tensor Fields. Proc. Int. Conf. on Information Visualization Theory and Applications, 2012, 670-677.

[27] Overhage JM, Grannis SJ, McDonald CJ. A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions. Am J Public Health. 2008 Feb;98(2):344-50. PubMed PMID: 18172157.

[28] Stephen G. Kobourov. Spring Embedders and Force Directed Graph Drawing Algorithms. arXiv: 1201.3011.

[29] B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–11

[30] S. Osinski, D Weiss. A concept-driven algorithm for clustering search results - Intelligent Systems, IEEE Intelligent Systems, May/June, 2005. pp. 48-54.

[31] Nathan Gossett, Baoquan Chen. Paint Inspired Color Mixing and Compositing for Visualization. IEEE Symposium on Information Visualization 2004. 113-117.

[32] Ken Perlin. An image synthesizer. In Proceedings of SIGGRAPH85, pages 287–296. ACM Press, 1985.

[33] Rosenfeld, A. and A.C. Kak (1982). Digital Picture Processing. Academic Press, New York.

[34] Hoschek, J., (1988), “Spline Approximation of Offset Curves,” Computer Aided Geometric Design, Vol. 5, pp. 33–40.

Health-Terrain: Visualizing Large Scale Health Data

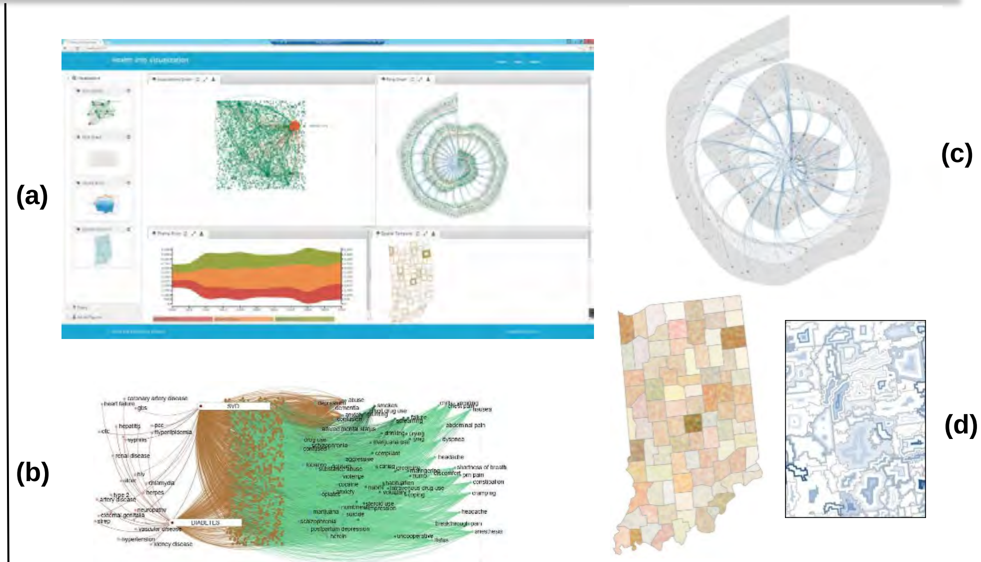
PI: Shiaofen Fang **Org:** Indiana University **Requested Amount: $670,646**

Study/Product Aim(s)

•Build a framework for concept space for pilot healthcare applications.
•Extract attributes of concepts from raw health data using data mining and knowledge extraction.
•Develop and implement visualization algorithms for Health-Terrain visualization, and develop a prototype system.
•Develop application case studies using the prototype system.

Approach and Military Relevance

Raw health data will first be transformed to an information-rich Concept Space, where attribute values and association relations are extracted from patient data, and visualized using Health-Terrains. Health-Terrain is a terrain surface based scalable visual representation for large multi-dimensional data. It is an ideal approach for military EHRS as it is very effective in revealing trends, patterns, and abnormalities with simple heads-up displays to support real time decision making. It also provides a way to unify heterogeneous data into an intuitive and comprehensive visualization to facilitate interoperability among multiple military HER systems. Pilot cases will be studies using the nation's largest and longest tenured health information exchange through Regenstrief Institute.



(a) System interface; (b) Association graph; (c) Spiral Them Plot; (d) Texturization.

Timeline and Cost

Activities CY	13	14	15
Concept space def. (aim 1)			
Algorithm design (aim 1,2)			
System dev. (aim 2,3)			
System Testing (aim 2,3)			
Pilot appl. (aim 4)			
Estimated Budget ($k)	$100	$480	$90

Updated: April 3, 2014

Projected Goals/Milestones:

1. **3/7/2013 – 7/31/2013. Concept Space Definition:** data processing, definition of the concept set, including their attributes and functionalities based on the pilot applications.
2. **6/1/2013 – 12/31/2013. Algorithms Design:** completing the algorithm design phase for visualization and data mining.
3. **9/1//2013 – 12/31/2014. System Development:** finishing software development of the prototype system.
4. **9/1/2014 – 9/31/2015. System Testing:** completing functionality tests for all components of the system.
5. **9/1/2014 – 9/31/2015. Pilot Applications:** Testing the prototype system on pilot applications using the Regenstrief database.